

BTP Project Update-4

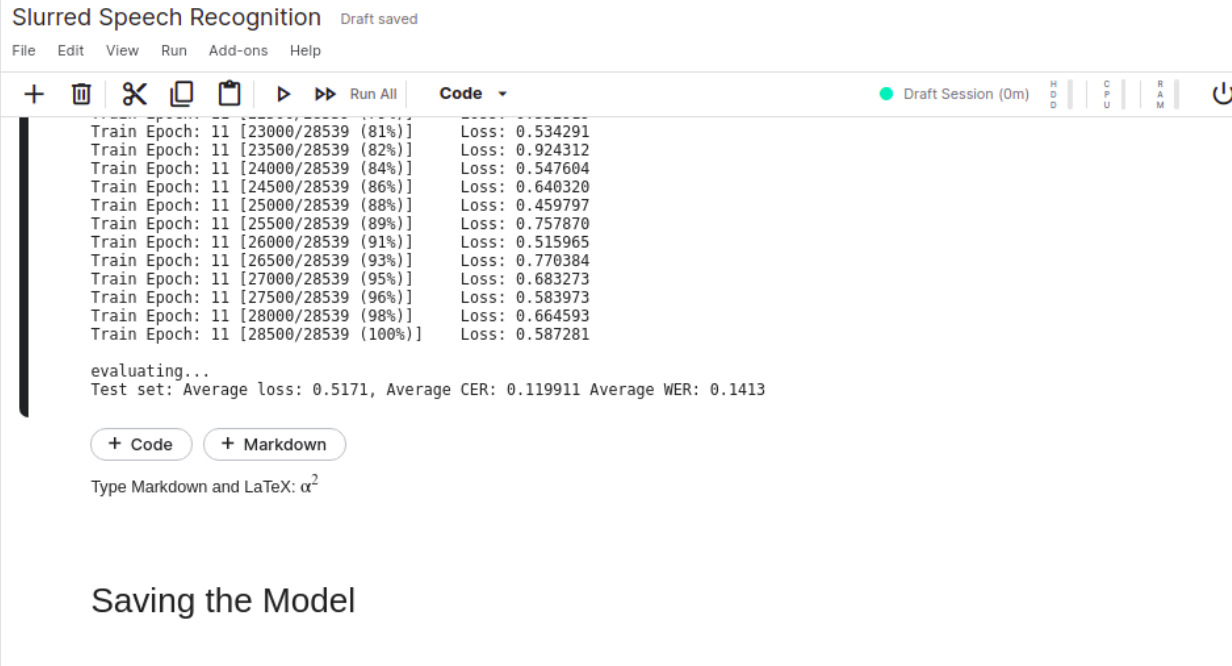
High-level Progress Overview

Last week, we completed the training of our end-to-end ASR deep learning model prior to transfer learning. We were also able to save the best fit model with a **Word error rate** of 0.17 and a **Character error rate** of 0.14 with epoch=5.

This week we performed mainly two tasks.

- Personalized Speech Dataset collection and preparation for fine-tuning our base ASR model.
- Improving the performance of our base model.

Our Base model WER now is 0.14 and CER now is 0.11.



The screenshot shows a Jupyter Notebook titled "Slurred Speech Recognition" with a "Draft saved" status. The interface includes a menu bar (File, Edit, View, Run, Add-ons, Help) and a toolbar with icons for adding, deleting, and running code cells. The main content area displays a table of training progress for 11 epochs. Each row shows the epoch number, the current step out of 28539, the percentage of training completed, and the loss value. The loss decreases from 0.534291 in epoch 1 to 0.587281 in epoch 11. Below the table, the text "evaluating..." is followed by the test set performance metrics: Average loss: 0.5171, Average CER: 0.119911, and Average WER: 0.1413. At the bottom of the notebook, there are buttons for "+ Code" and "+ Markdown", and a prompt "Type Markdown and LaTeX: α^2 ".

Train Epoch:	Step	Percentage	Loss
11	23000/28539	(81%)	0.534291
11	23500/28539	(82%)	0.924312
11	24000/28539	(84%)	0.547604
11	24500/28539	(86%)	0.640320
11	25000/28539	(88%)	0.459797
11	25500/28539	(89%)	0.757870
11	26000/28539	(91%)	0.515965
11	26500/28539	(93%)	0.770384
11	27000/28539	(95%)	0.683273
11	27500/28539	(96%)	0.583973
11	28000/28539	(98%)	0.664593
11	28500/28539	(100%)	0.587281

evaluating...
Test set: Average loss: 0.5171, Average CER: 0.119911 Average WER: 0.1413

+ Code + Markdown

Type Markdown and LaTeX: α^2

Saving the Model

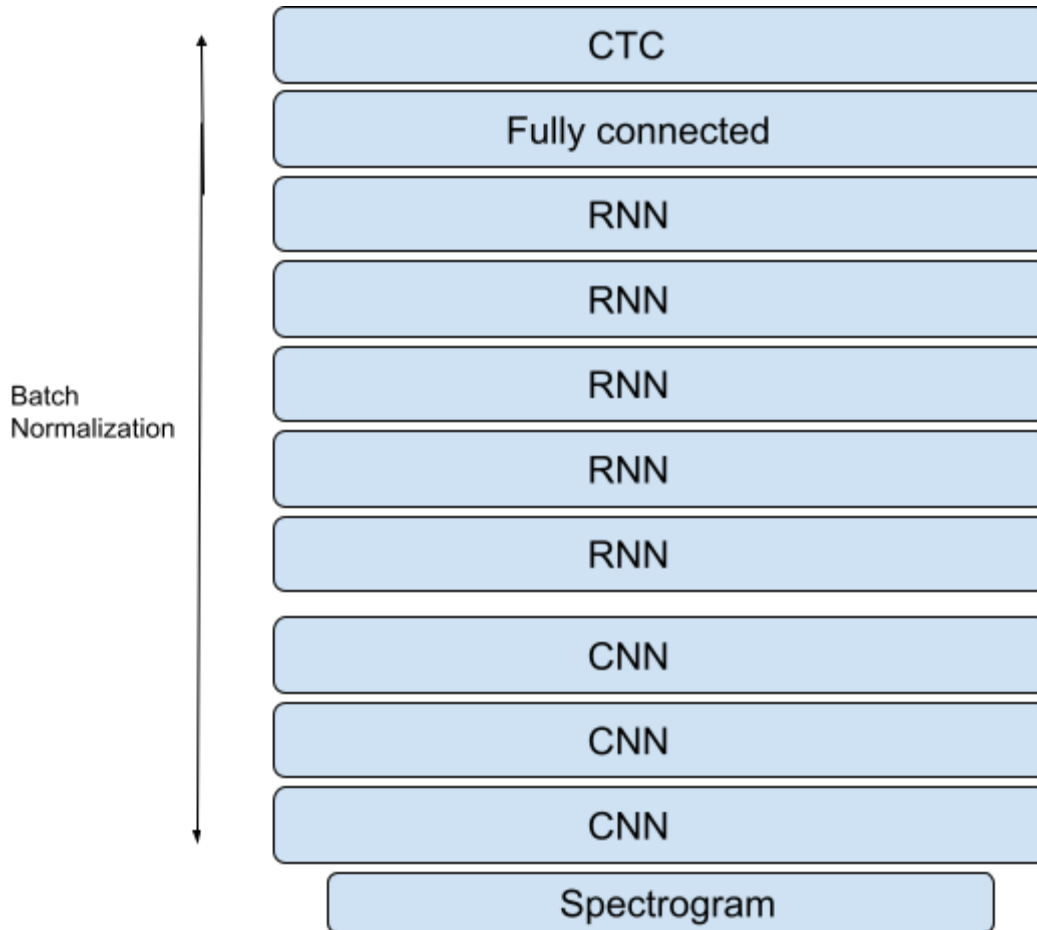
Low level Progress Overview

Steps performed till now:

- Data preprocessing
- Raw audio data augmentation
- Generation of Mel spectrograms
- MFCC
- Neural network architecture building
- Model Training
- Model Evaluation
- Saving Best fit Model
- Built Speech dataset Collection API using Django

Architecture of our model:

The architecture of our model will use Convolutional neural network(CNN) as well as Recurrent neural network(RNN) packed with CTC(Connectionist temporal classification) function that will calculate the WER(Word error rate) for our model. Below is the architecture of our model before transfer learning.



Model Performance:

- Final Epoch Average Loss: 0.61
- Final Epoch Average CER: 0.11
- Final Epoch Average WER: 0.14

Testing our loaded base model

```
[ ] test(loaded_model, device, test_loader, criterion, 1, iter_meter)
```

```
evaluating...  
Test set: Average loss: 0.5171, Average CER: 0.119907 Average WER: 0.1413
```

Snapshot of the Dataset collection Interface:

Speech Dataset Collection App

HomeAboutDropdown

Search

Search

ASR Dataset Collection API

This is a Web simple interface build with django used for providing support for personalised speech dataset collection used in our current project of slurred speech recognition.

Learn more

100 Transcripts left

SI NO	Transcript	Audio
0	apples and oranges are fruits not vegetables	<div>Choose File</div> No file chosen <div>Submit</div>
1	interstellar is a good science fiction movie	<div>Choose File</div> No file chosen <div>Submit</div>
2	the salt in the ocean reflects the light from the sun	<div>Choose File</div> No file chosen <div>Submit</div>
3	i always get very cold in the winter	<div>Choose File</div> No file chosen <div>Submit</div>

Django administration

WELCOME DELLVIEW SITE / CHANGE PASSWORD / LOG OUT

HomeAppDatasets

APP

DatasetsAdd

AUTHENTICATION AND AUTHORIZATION

GroupsAdd

UsersAdd

Select dataset to change

ADD DATASET

Action:Go0 of 100 selected

☐ DATASET

☐ Peanut butter is made by crushing peanuts and adding oils

☐ It is a beautiful and sunny day

☐ Bananas are yellow and not green

☐ Star Wars is an imaginary world that has multi-levels of races and planets

☐ They need to maintain law and order in the country

☐ My favorite pizza is made with the perfect ratio of crust to sauce

☐ I would prefer if cell phones did not change sizes

☐ My hair is a purple color but i wished it was black or blue or grey

☐ I had all my toys piled on one side

☐ I practice phone calls so I do not mess up

☐ I also like to eat cakes more than ples

☐ I sent flowers for the first time today

☐ My favorite animal has to be a dog

☐ I want to open a coffee shop

☐ Spanish has more letters compared to English

☐ I got a toy car when I was four years old

☐ Choosing avocados is very hard because quality matters a lot

☐ My family always takes care of me no matter what

☐ There are good and positive things in everyone

Link to web APP:

1. <https://mmig.github.io/speech-to-flac/>
2. <http://speech-collection.herokuapp.com/index/>

Next goals to perform:

1. Applying transfer learning paradigm
2. Building text to speech model

Future plan of action:

Since we are done with end to end ASR model training and built a solution for gathering impaired speech dataset , we can now focus on transfer learning. After that we will implement the next phase of our project i.e building text to speech model.

References:

1. <https://arxiv.org/pdf/1512.02595v1.pdf>
2. <https://arxiv.org/pdf/1412.5567.pdf>
3. <https://www.biometricupdate.com/201906/google-building-impaired-speech-dataset-for-speech-recognition-inclusivity>
4. <https://ai.googleblog.com/2019/08/project-euphonias-personalized-speech.html>