CS 753: Assignment #2

Instructor: Preethi Jyothi Email: pjyothi@cse.iitb.ac.in

April 4, 2022

Instructions: This assignment is due on or before 11.59 pm on April 17, 2022. No grace period will be granted for this assignment. The submission portal on Moodle will be closed after 11.59 pm on April 17.

- This is a group assignment. You will be starting with a pretrained end-to-end ASR model and adapting it to small amounts of labeled speech in an Indian language. This is a use-case that is popularly adopted in both industry and academic settings for speech and NLP applications.
- Click here for the structure of your final submission directory. It is **very important that you do not deviate from the specified structure.** Deviations from the specified structure will be penalized. All the files that need to be submitted are highlighed in red below; all the submitted files will be within the parent directory `submission/`. Compress your submission directory using the command: `tar -cvzf submission.tgz submission` and upload `submission.tgz` to Moodle.

Adapting an End-to-End English ASR System to Recognize Marathi

Pretrained end-to-end ASR models for English are easily available in a variety of toolkits. For this assignment, we will use the Coqui Toolkit.

Task 1: Finetune to Recognize Marathi [5 points]

Here is a Python notebook from Coqui detailing how to start with a pretrained model and finetune it using a single speech file in Russian: Transfer learning.

Make minor edits to this notebook to adapt the pretrained model to do ASR for Marathi instead. Download the Marathi data here. This archive file, once expanded, consists of ten wav files within `data/wavs/train`, five wav files within `data/wavs/test` and the transcripts are listed in `data/marathi.tsv`. (Note: You will have to create your own alphabet file for Marathi.)

What to submit: With the default model settings, what is the WER obtained if you test on the training instances in `data.tgz`? Write down your answer in a text file `task1/wer-train.txt`. What is the WER of this model on the test utterances in `data/wavs/test`? Write down your answer in a text file `task1/wer-test-baseline.txt`. Tweak the hyperparameters in any meaningful way you like to improve performance on the test set and write down the best WER you obtain in `task1/wer-test-best.txt`. Submit a link to your notebook in `task1/notebook.txt` that we can run on Colab to reproduce your best results. This should run end-to-end to produce the final test results and all the hyperparameters should be set to the best values. You will receive full points for this question if your test WER with the tweaked model is lower than the WER we get for the test set with the default hyperparameter settings.

We will also evaluate your model on a blind test set and share a leaderboard. The N top-scoring performers on the leaderboard, based on WERs on the blind test set, will gain extra credit points.

Useful read: Refer to the following paper for hints and pointers regarding which hyperparameters might be effective to tune.

Task 2: Multilingual Training [10 points]

Multilingual models are a very popular paradigm when developing speech and language technologies for new languages. Download the multilingual data available here. Expanding this .tgz file will give you a `train` directory consisting of four directories named `tamil`, `telugu`, `malayalam` and `kannada` with 10 wav files each. `train.tsv` contains all the training transcripts. The folder `test` consists of 10 wav files in Kannada with their corresponding transcripts in `test.tsv`.

Your aim is to build a Kannada ASR system using the multilingual training data in any way. E.g., you may choose to completely ignore the Malayalam and Tamil wav files and only use the Telugu and Kannada training files.

What to submit: Implement the best training schedule/model you have identified using Coqui and submit a link to your notebook in `task2/notebook.txt` that we can run on Colab to reproduce your results. What is the WER obtained with the model implemented in your notebook when tested on the instances in `ml-data/test/test.tsv`? Write down your answer in a text file `task2/wer-test.txt`. You will receive full points for this question if the test WER with your model is lower than the WER we get with finetuning on only five training instances of Kannada, one from each speaker in `ml-data/train/kannada`.

We will also evaluate your model on a blind test set and share a leaderboard. The N top-scoring performers on the leaderboard, based on WERs on the blind test set, will gain extra credit points.

You are allowed to use external resources (e.g., text to train an LM, raw audio, etc.). These should be linked within the notebook and publicly available so that we can download it during grading.

References: Some useful pointers to things that you could try out to improve performance:

- You could train an LM on Kannada text and link it to a beam search decoder. Refer to the following page on how to use an external LM within Coqui.
- You could try data augmentation on the existing training samples. See here for more pointers on sample augmentations.