COMPSCI304 Spring 2023 Project 7 Due date: March 9th 9am Beijing time

Problem 1 (15 points):

In this problem, you need to use the KALDI toolkits to train and test a GMM-HMM based large vocabulary continuous speech recognition system.

We provide a KALDI example called aishell1. You need to log in your group's own VM to run the example where the dataset and script are provided. (please check quickstart.md)

When the training process of the aishell example completes, you need to report the final experimental results, such as the one we showed you in the lecture, RESULT.txt.

You need to write a REPORT to explain the main steps in the aishell example script (run.sh and other major tools it uses), e.g. the data preparation, feature extraction, acoustic model training and the testing phases, etc. (please check the rubrics file)

As for data preparation, local folder contains the code related to data preparation, and you need to explain the codes in details.

As for model training, the introduction to the model training script does not need to go into specifics, whereas you need to pay efforts in explaining the concepts related to the current training step.

As for model testing, you need to describe in detail the evaluation criteria for the test section.

Bonus Question: (3 points)

We provide another three datasets, which are extracted from three open source datasets. you need to complete the data preparation steps, integrate the three datasets and then divide the merged set into train set and test set. And refer to the aishell script to build your own training steps, and test the model performance on this test set, and also test on the aishell test set.

All the provided datasets have their original audio files and transcribed text. The speaker information is included in the file directory.

Please note that there should be no overlap between the speakers in the train set and test set.

You may need a word segmentation tool. If so, jieba is recommended.

https://github.com/fxsjy/jieba

Appendix:

1. Report Example:

….

local/aishell_data_prep.sh $data/data_aishell/wav $data/data_aishell/transcript || exit 1;

This step is to generate wav.scp,text,utt2spk,spk2utt for describing data information. ….

The detail meaning of these files are:

wav.scp: ….

text: ….

utt2spk: ….

spk2utt: ….

….

2. We provide a markdown file called “QuickStart.md” helping you quickly start the aishell script.
3. We provide a markdown file called “BonusQuetion.md” to introduce the extra three datasets in detail.