**CS 768 Project Problem**

**Project Title: Pre-training on High-Resource Speech Recognition for Low Research Speech-to-Text Translation**

**Introduction Section**

**Task / Research Question Description**

We aim at solving task of translating speech (audio) into text for low resource languages. Given audio data in a source language, we aim at directly translation the audio to a target language into text.

**Motivation & Limitations of existing work**

The main motivation comes from the fact that languages with low resources (limited data) do not directly benefit from recent advances in speech transition due to limited access to a corpus that can be used from training a model. We aim extending speech to text translation models in cases where we do not have enough samples in the corpus to train a model for a given language pair. There has been prior work in trying to solve the same problems. The literature we surveyed use sequence to sequence architectures with or without attention. We would like to approach this task by using the transformer architecture which has showed superior performance in various natural language processing tasks

**Proposed Approach**

We aim at using pre-training mechanism where we leverage data from a high speech recognition task to pretrain the model before finetuning the learned model on a low resource task. We want to explore the use of the transformer architecture for this task.

**Likely challenges and mitigations**

Speech translation is complex tasks. We plan on using high level frameworks in our development process. We plan to be able to get the expected behavior from these frameworks. In the event we are saucerful to bend these frameworks to our needs, we may need to reimplement different components from scratch. We also plan to use the hopper/argo cluster from George Mason University. Depending on the resources available, it may be difficult to run multiple experiments faster as we iterate through the version of our project. We can prioritize experiments have baseline results from each phase of the experiments before running more experiments.

**Related Work Section**

[1] and [3] have suggested speech to text datasets and provide baseline results but did not use pretraining with speech data from high resource languages. [2] and [4] use pretraining in their training. However, we want to explore different pretraining methods and the impact of pretraining of related and unrelated languages.

**Experiments section**

For the dataset, we plan to use the Multilingual TEDx which is a multilingual corpus in the development process because its publicly available. We also choose this dataset because it’s been used in previous tasks by other researchers [] in giving baseline results. However, the Multilingual TEDx consists of high resource languages. We plan to simulate the high low resource language by sampling a submit of samples from a given language pair (i.e., French to Spanish).

We will implement the baselines from [1]. [1] used the CoVoST 2 dataset. We plan to run their model on the Multilingual TEDx dataset and compare their results with the baseline on the Multilingual TEDx [2] baselines.

We will use the fairseq framework [2] which is built on top of Pytroch. Pytroch is one of the most popular frameworks for deep learning application. Fairseq is one of the most popular frameworks particularly for speech various NLP translation tasks including speech translation.

Timeline:

|  |  |
| --- | --- |
| Week | Task |
| 10/03 – 10/14 | Reproduce the baseline from [1] |
| 10/15 – 10/21 | Reproduce the baseline from [1] |
| 10/22 – 10/28 | Prepare the Multilingual TEDx [2] corpus to run on the reproduced model |
| 10/29 – 11/03 | Running the model on the prepared corpus |

10/7 - 11/03

Reference

[1] CoVoST 2 and Massively Multilingual Speech-to-Text Translation

[2] The Multilingual TEDx Corpus for Speech Recognition and Translation

[3] Leveraging translations for speech transcription in low-resource settings

[4] Pre-training on high-resource speech recognition improves low-resource speech-to-text translation