**Report - Homework 2**

Github Repository: https://github.com/Anirudh-Kannan/Multilingual-Speech-Recognition

| Name | Andrew ID |
|---|---|
| Anirudh Kannan | akannan3 |
| Sathyanarayanan Ramamoorthy | sramamoo |

In this homework, we attempted to train ASR models for Guarani, exploring individual as well as multilingual ASR Models.

**Sub Task 1**

In subtask 1, we created the ASR model using fairseq.

- *Preprocessing*: We made code modifications to the existing preprocessing file for librispeech. This includes renaming the splits, changing the paths to the source files as well as reshaping the waveform array to make it compatible for fbank feature extraction.
- *Modeling:* We used the s2t_transformer with label smoothed cross entropy loss, adam optimizer, learning rate of 2e-3, inverse_sqrt scheduler and a max epochs limit of 121.
- *Results:* We achieved a **WER of 98.58** with beam=5.

**Sub Task 2**

In subtask 2, the focus was on fine-tuning the XLSR, which is the multilingual variant of the wav2vec2 model. The following were some common settings used in the experiments

- Removed special characters like ., ?, ! during preprocessing.
- To optimize the model's performance, we employed the CTC (Connectionist Temporal Classification) Loss function. This choice is primarily because the CTC Loss function is well-suited for sequence-to-sequence tasks like speech recognition and enables one to one correspondence.
- To monitor the model's progress the Word Error Rate (WER) was calculated and validated every 25 training steps.
- Model was trained with an initial learning rate of 3e-4 in batches of 16.
- We employed a sampling rate of 16,000 in the feature extractor.

In our efforts to improve the Word Error Rate (WER) for Guarani and Quechua, we conducted four distinct experiments:

- **Experiment 1**: In this initial experiment, we fine-tuned the XLSR model exclusively on the Guarani language dataset. The model's performance was evaluated on the test split of Guarani from the commonvoice 13 dataset. After **50 epochs** of training, the WER was observed to be **0.48**.

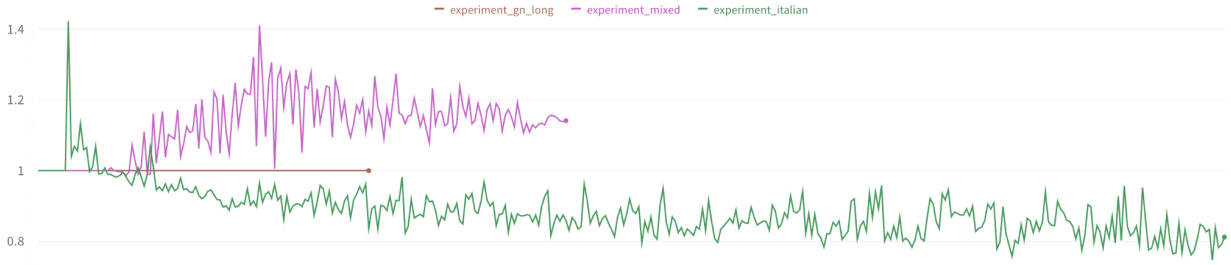
- **Experiment 2**: For the second experiment, we attempted to enhance the dataset by incorporating a similar language for fine-tuning, and we explored cross-lingual transfer learning. Italian, with its linguistic similarities to Guarani, was chosen for this purpose. While there are not many linguistically similar languages to Guarani, we considered

Italian due to historical and geographical influences, some linguistic commonalities, and the Romance language factor. However, due to the substantial size of the Italian dataset, even after 22 hours of (incomplete) training, we could only achieve a WER of **0.88**

- **Experiment 3**: In another experiment, we sought to enhance Guarani's Word Error Rate (WER) by utilizing a combination of Portuguese and Italian datasets for cross-lingual transfer learning. While we employed only 10 percent of each of these datasets to avoid overwhelming the training process, the endeavor presented challenges. The training proved to be time-consuming, and the WER did not exhibit signs of convergence. We achieved a WER of **0.99**.

- **Experiment 4**: In addressing the limited dataset for Quechua, which contained just one example in the training set, we used this single example as a basis for evaluating the model trained in Experiment 1. This approach yielded a WER of **1.0** for Quechua. This was primarily due to the fact that Quechua had only one sample and we used this sample for evaluation. Since the model was not trained on any samples from Quechua, it performed poorly in a zero shot setting.

WER variation plot for Experiments 2, 3 and 4.




**Conclusions**

The efforts to develop ASR models for Guarani showed mixed results. While the fairseq-based s2t_transformer ASR model was not effective, with a **WER of 98.58**, fine-tuning the XLSR model on Guarani alone showed promise, **reducing the WER to 0.47647**. Cross-linguistic transfer learning, using Italian and a combination of Italian and Portuguese, did not yield improvements and, in the latter case, resulted in performance degradation with the model diverging. Finally, the experiment involving Quechua highlighted the challenges of working with extremely limited data.

These results emphasize the challenges in low-resource ASR model training, where data scarcity and language complexity can significantly hinder performance. Future work could involve gathering more comprehensive datasets, exploring other potentially similar languages for transfer learning, and experimenting with different model architectures similar to Wav2Vec2 and variants or better training strategies to improve the ASR performance for low-resource languages like Guarani and Quechua.

Our WandB Plots can be found here:

- https://api.wandb.ai/links/learntorace/8gvwhzef
- https://api.wandb.ai/links/sathyaram/xg56h300