Multilingual Speech Recognition

Group Members: Kshitish Ghate Vishwa Shah

The goal of this task is to construct an Automatic Speech Recognition system for any language, in this assignment we choose Guarani as our subject language. Guarani is a South American language that belongs to the Tupi–Guarani family of the Tupian languages. It is one of the official languages of Paraguay and also spoken by communities in parts of the neighboring countries Argentina, Bolivia and Brazil.

Sub-Task 1:

For sub-task1, we choose the Fairseq[2] framework. Since Fairseq does not support CommonVoice dataset directly, we make several changes in the preprocessing part (refer to prep_librispeech_data.py in the code.

1. We first define the CommonVoice dataset class which loads training samples in accordance with CommonVoice's data path configuration. When obtaining each training instance, this class extracts the waveform and sample rate from raw audio files using the torchaudio library.
2. Extract FBank Features - After obtaining the waveform we extract fbank features that uses Mel Filter bank to return Mel-Frequency Cepstral Coefficients (MFCC) - this is used as feature input to our transformer.
3. For tokenization we use the unigram level tokenization to handle the distinct character and lemmatization of Guarani.

Below are the results for training the S2TTransformerEncoder model using the CommonVoice 15.0 dataset for Guarani (Train and Test) and Cross Entropy based loss (with label smoothing)

Model	Loss	Val WER(%)	Test WER (%)
XLS-R-Guarani	9.715	138.83	185.00

Since the Guarani dataset contains only about 1.5k training instances after preprocessing, it is incapable of performing ASR as it obtains WER of > 100 %

Sub-Task 2:

For sub-task 2, we explored advanced techniques like finetuning and multilingual finetuning of pre-trained models to build ASR systems for the Guarani language. We experimented with two popular pre-trained models designed for speech recognition:

1. XLS-R 53: This model utilizes a transformer architecture and is pre-trained on hundreds of thousands of hours of speech data across 53 languages in a self-supervised fashion, similar to how BERT is pre-trained on text. After pre-training, only a simple linear layer needs to be

fine-tuned on labeled data to adapt XLS-R for downstream speech tasks like ASR. The architecture and massive multi-lingual pre-training enables XLS-R to learn powerful contextual speech representations that transfer well to low-resource languages like Guarani.

2. Whisper: This model follows a similar paradigm but is different in that it is pre-trainined on 680,000 hours of labeled speech data, including 117,000 hours of multi-lingual data spanning 96 languages. This allows Whisper to acquire extensive knowledge of multilingual speech during pre-training. The model can then be fine-tuned on small labeled datasets to adapt to specific languages and even low-resource languages. Whisper demonstrates an impressive ability to generalize across languages and domains thanks to this abundant labeled pre-training.

Both XLS-R and Whisper leverage self-supervised pre-training on large labeled speech datasets to learn powerful contextual speech representations. Fine-tuning them on small labeled Guarani speech data then allows us to build high-quality ASR systems even for low-resource languages. We experimented with both models to evaluate their performance on the Guarani language.

We use Spanish as our source language as Spanish and Guarani share geographic similarity and also constitute lexically similar components, making it a suitable language for positive transfer. For multilingual fine-tuning, we used a XLS-R model fine-tuned on Spanish splits from Common Voice which is available on huggingface[1], and fine-tuned it further on Guarani. We report the results for all three models below, As the latest CommonVoice version of datasets on HuggingFace is 13.0 we use the same for SubTask-2 Train and test:

Model	Test Loss	WER (%)
XLS-R-Guarani	0.4991	52.98
XLS-R-Spanish-Guarani	0.3713	35.70
Whisper-Guarani	0.6625	54.79

As XLS-R uses the Connectionist Temporal Loss during finetuning - it helps the model attain much better alignment which is beneficial for a task such as ASR, achieving much better performance than model in subtask1. The multilingual XLS-R model finetuned on both Spanish and Guarani achieved the lowest test loss of 0.3713 and the best WER of 35.70% on the evaluation set. This demonstrates the benefits of leveraging transfer learning from high-resource languages like Spanish when building ASR for low-resource languages like Guarani. The XLS-R model trained on Guarani alone achieved higher test loss and WER, indicating multilingual transfer learning improves performance. Whisper did not perform as well as XLS-R, achieving much higher test loss and WER when trained on Guarani alone. Overall, the Spanish XLS-R model achieved the best ASR performance with finetuning on our Guarani speech recognition task.

[1] https://huggingface.co/facebook/wav2vec2-large-xlsr-53-spanish

[2] https://github.com/facebookresearch/fairseq