Adding custom languages to NeMo ASR models

Dataset requirement:

JSON file for train & validation datasets.

https://docs.nvidia.com/nemo-framework/user-guide/latest/nemotoolkit/asr/datasets.html#preparing-custom-asr-data

The audio_filepath field should provide an absolute path to the .wav file corresponding to the utterance. The text field should contain the full transcript for the utterance, and the duration field should reflect the duration of the utterance in seconds.

{"audio_filepath": "/path/to/audio1.wav", "text": "the transcription of the utterance", "duration": 23.147}

{"audio_filepath": "/path/to/audio2.wav", "text": "second transcription of the utterance", "duration": 27.147}

....

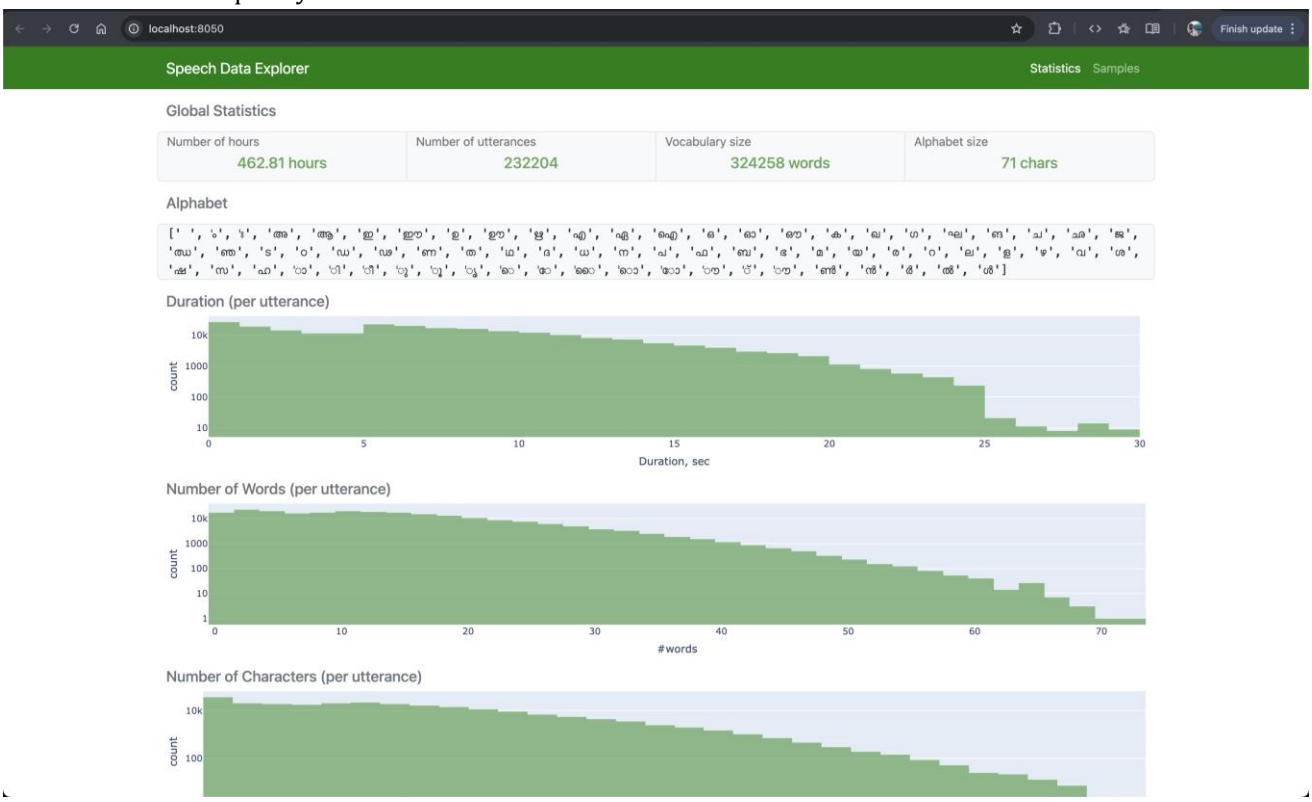
Docker:

Container image: nvcr.io/nvidia/pytorch:25.01-py3

Speech Data Explorer:

https://docs.nvidia.com/nemo-framework/user-guide/latest/nemotoolkit/tools/speech_data_explorer.html

This can be used to quickly understand various attributes of the dataset



Training:

1. **This notebook is used to create updated vocabulary using text corpus, the pre-existing Es(Espanol) can be updated to any regional language short code & respective training/validation paths as well.**
    a. https://github.com/weiqingw4ng/NeMo/blob/fixing_multilangASR_tutorial/tutorials/asr/Multilang_ASR.ipynb
    b. **Streaming compatible models https://docs.nvidia.com/nim/riva/asr/latest/support-matrix.html#supported-models**
2. **Save a variant of any pre-trained NeMo model with the new tokenizer as shown in notebook, omit En tokenizer if not needed for mono-lingual model**
3. **Start training using linked defaults depending on the model size chosen-**
    a. https://github.com/NVIDIA/NeMo/blob/main/examples/asr/conf/fastconformer/fast-conformer_ctc_bpe.yaml
    b. Wandb account API key to view the same on wandb.ai portal