Spoken language recognition from audio: Existing model assessments and new experiments

Adhilsha and Aritra Mukhopadhyay (Group 2)

Dataset: Self-made Dataset consisting of copyright-free audio clips in Bengali and Malayalam.
Objective: To train our model with our labeled data and expect it to recognize the language of an unseen audio clip. We will start with two Indian Languages: Bengali and Malayalam.
Relevant Papers:

- J. Valk & T. Alumäe. VoxLingua107: a Dataset for Spoken Language Recognition
- A. Radford et al. Robust Speech Recognition via Large-Scale Weak Supervision
- A. Babu et al. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale
- H. N. Krabbenhöft & E. Barth. TEVR: Improving Speech Recognition by Token Entropy Variance Reduction

Work Division

- **Aritra:** Downloading Bengali audio clips, cleaning data, implementing the models, and analyzing the results, Reading papers.
- **Adhil:** Downloading Malayalam audio clips, tuning hyperparameters, and analyzing the results. Reading papers. Documenting the project.

Midway Plans:

There are two major parts to midway plans: Data preparation and Implementation of some preliminary model:

1. **Data Preparation** includes converting all audio to '.wav' format => Concatenating the clips => dividing them into comparable length sequences => basic preprocessing and extraction of features (using methods like **MFCC, BFCC, PLP** etc. or even preprocessing models like **wav2vec2.0** from fairseq can be tried) based on different papers.
2. Learning about different **models** relevant to audio recognition and implementing some preliminary models from them.

Further Plans:

- Analysing preliminary model performance and tuning their hyperparameters according to our dataset.
- Augmenting dataset (if needed) by adding noise, changing pitch, etc
- Study the model's behaviour on different adversaries like letting it predict an unseen language.
- Possibly update the model based on insights obtained from these experiments.

Expected Results:

The baseline would be to successfully extract different features and implement some **CNN or Transformer** based models if not more. Obtain insights related to language models and features which are prioritized over others. Possibly improve the accuracy of our model with this information at hand.