MAIS 202 – Deliverable 1 (Language Detection Model)

Brief Description:

Welcome to our Language Detection Model! This project aims to automatically identify the language of a given text or voice note using signal processing, making it a versatile tool for applications such as content filtering, language-specific analysis, and more.

Note that our choice of datasets and algorithms are subject to change and will be finalized as we progress further into our project.

Dataset chosen:

We have temporarily decided on https://huggingface.co/datasets/mozilla-foundation/common_voice_13_0 as our choice of dataset. We chose this because it consists of unique MP3 recordings (as well as text files) with language classifications, which is what we were looking for.

Additionally, many of the 27141 recorded hours in the dataset also include demographic metadata like age, sex, and accent that may help improve the accuracy. Although there are 100+ languages in the dataset, we will be focusing on well-known languages such as English, Mandarin, French, Italian, Arabic, Hindi/Urdu, Russian, Spanish, etc.

Methodology:

Data Preprocessing: Since the dataset consists of audio files, the preprocessing step will involve converting these MP3 files into spectrograms or MFCCs, which are standard tools for working with audio data. The demographic metadata, such as age, gender, and accent will be utilized as additional features to enhance the model's performance.

Machine Learning Model: Our goal is to create a model that can identify the language from a given audio clip. A multiclass classification approach where each language represents a class would work well for our purposes. A CNN or CRNN seems like a suitable architecture for this task, as they're each often used for similar audio classification tasks (e.g. music genre classification). CNNs may be simpler and faster to train and build than a CRNN, but they have limited capacity to understand longer time-based patterns.

Evaluation Metric: Since it's a classification problem, the main metrics to consider are classification accuracy, precision (false positives), sensitivity (false negatives), and accuracy by language and given features.

Application:

The goal for now is to implement this language identification model into a simple web app with a GUI to record and upload audio and then display the predicted language.