Spoken Language Identification from Audio for Uncommon Languages

This project involves identification of the language spoken in audio. The idea is to first translate the audio data into text using a speech recognizer. As the next step, the recognized text will be processed to identify and tag the corresponding language. This project will also explore other methods to directly identify language from the audio, instead of having an intermediate audio to text conversion. This will involve preprocessing and feature extraction steps from the audio data and supplying these features as input to the model. The project will focus on less resourced languages on YouTube. Uncommon Indian languages from https://commonvoice.mozilla.org/en/languages will be used as the datasets to train the language identification model. To evaluate the model, audios downloaded from YouTube will be used as evaluation/test dataset. This project will be useful to tag uncommon languages spoken in YouTube videos. Tagging these uncommon languages will help in expanding the number of languages for which subtitles can be submitted on YouTube. Although the exact architecture of the model has not been designed yet, a suitable type of neural network will be an appropriate choice.

Bibliography:

1. Christian Bartz, Tom Herold , Haojin Yang and Christoph Meinel, Language Identification Using Deep Convolutional Recurrent Neural Networks - https://arxiv.org/pdf/1708.04811.pdf
2. Julien Boussard, Andrew Deveau, Justin Pyron, Methods for Spoken Language Identification - http://cs229.stanford.edu/proj2017/final-reports/5239784.pdf
3. Priyank Mathur, Arkajyoti Misra, Emrah Budur Garanti, Language Identification from Text Documents - http://cs229.stanford.edu/proj2015/324_report.pdf
4. Mounika K V, Sivanand Achanta, Lakshmi H R, Suryakanth V Gangashetty, and Anil Kumar Vuppala, An Investigation of Deep Neural Network Architectures for Language Recognition in Indian Languages - https://www.semanticscholar.org/paper/An-Investigation-of-Deep-Neural-Network-for-in-MounikaK.-Achanta/5f6ffd39e74a66492cfb34b62a21e91d08332e35
5. Krishna D N, Ankita Patil, M.S.P Raj,Sai Prasad H S, Prabhu Aashish Garapati, IDENTIFICATION OF INDIAN LANGUAGES USING GHOST-VLAD POOLING - https://arxiv.org/pdf/2002.01664.pdf
6. Shauna Revay and Matthew Teschke, Multiclass Language Identification using Deep Learning on Spectral Images of Audio Signals - https://arxiv.org/pdf/1905.04348.pdf