Spoken Language Identification (SLID)

Goal: Given an English or Mandarin Chinese sentence spoken by the user, the program will be able to identify the language spoken. (As of yet, I am unsure how many languages I will include and how closely overlapped these languages will be. I will start with English and Mandarin Chinese, then depending on the success of that experiment, I will add more languages.)

Possible Datasets

EN, DE, ES, Common Voice

Each language included will ideally have at least 50 hours of validated audio samples and 100 different voices with a good distribution of female and male voices. For languages like English, we will take care to choose a dataset that includes native speakers from various countries with some variation in regional accents.

Methodology

Data Preprocessing:

1. Audio processing using Python → Librosa or torchaudio (less good than librosa, but better integrated with Pytorch)
    - This will allow us to load the files, visualize the sound wave and play the audio
2. Start by cleaning the audio samples to get rid of blank stretches of time when there are no words spoken and general "noise", align speech signals, adjust the audio files based on loudness... → SpeechBrain
    - Several research papers have mentioned getting rid of the silent portions to prevent them from affecting the classification results
3. Turn the audio files into Mel spectrograms that can be easily fed into CNN-based models that are generally built for images
    - Fourier Transform applied to each interval of the audio and then recombined together to make a frequency vs time graph; however, we usually use a Mel Scale on the y-axis
    - https://ketanhdoshi.github.io/Audio-Mel/
    - https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html
4. Feature extraction:

    The biggest challenge will be to extract and identify the features that distinguish one language from another (especially when the languages have significant overlaps, e.g. romance language). For this reason, we will start off using two languages that are very distinct and then attempt to include more languages. We will do the feature extraction using Mel Frequency Cepstral Coefficient (MFCC) which appears to be one of the most commonly used feature extraction method for SLID .

Machine learning model:

Table showing the accuracy of different models on various datasets

Based on this table, it appears that a Convolutional Neural Network showed the best results. The number of layers or the specific architecture of CNN used has not been decided.

Speech Command Recognition with Convolutional Neural Network

Evaluation Metric:

The performance of our model will be depicted using a Confusion matrix.

From a general overview of the literature, it seems that the various models have an accuracy ranging from 70% to 97% on average depending of course on the number of languages, the choice of languages used and the actual choice of the model. For this project, we are hoping to predict the language spoken with at least 75% accuracy (considering the fact that the audio recorded by the user will perhaps not have been processed to get rid of background noise and the possibility of accents, age group, gender... affecting the results).

Application:

What does the user input? How does the user provide inputs?

The user will input a voice recording of them saying a short phrase in the language of their choice (limited of course by which languages the model has been trained on). The audio can be recorded using the microphone on their computers and imputed directly from there.

What does the user receive as output, and how will the output be displayed?

The user will see the language the model predicts along with the percent confidence. May potentially add a rotating globe that shows all the regions around the world where the identified language is spoken for the aesthetics

Praat (phonetics)

SpeechBrain (features extraction, normalization)

https://huggingface.co/speechbrain/lang-id-voxlingua107-ecapa

Spoken Language Identification Using Deep Learning

Acoustic Phonetic Feature Extraction

https://www.mdpi.com/2076-3417/12/18/9181

Alternative to spoken language identification:

https://www.kaggle.com/datasets/zarajamshaid/language-identification-datasst