MULTILINGUAL CLASSIFICATION AND MODEL SELECTION

Marcos Paulo Pereira Moretti

Number: 9345363
Escola Politécnica da USP
Course: Master's degree
Domain: Computer Engineering
PSI5123 - Aprendizagem de Máquina de Sinais de Áudio e Voz (2022)
São Paulo - SP, Brazil

ABSTRACT

This paper presents a proposal for a project of a multilingual classifier based on audio recordings in six different languages. Data from OpenSLR will be sampled and balanced by gender and language, and used for training, testing and choosing the best among different models. The audio data will be prepared, transformed and modelled using different techniques. We will report the values of classification metrics for every model, and choose the best model based on this metrics. As a plus point to this project, we're going to try to identify relevant features for each language based on the speech.

***Index Terms*—** Multilingual, speech, modelling, multiclass classification, classification metrics.

1. INTRODUCTION

This paper presents a proposal for the development of many models aiming to classify audio recordings according to the language spoken by the speakers. Nowadays, there are many applications regarding to multilingual classification; for instance, Google Translator does a good job on this topic, being able to recommend the spoken language given an speech audio or typed text. However, as there may be beginners in the field of Audio and Speech Processing, it may be confusing why certain modelling methods may fit better to audio records than others. The main objective of this research is to do audio preprocessing, and then apply many kinds of modelling techniques regarding to multilingual classification, compare them, and select the best option given a specific dataset.

2. DATASETS

For this project, we choose free datasets produced by OpenSLR [1], in the compressed fashion. We will choose the datasets for six of the eight offered languages: Polish, Portuguese, Italian, Spanish, French and Dutch. This dataset is derived from read audiobooks from LibriVox.

We're going to filter the dataset in every language, so as to have at least 1 hour of speech for every language, as well as half of the dataset spoken by men and half by women. The reason for that is to have enough diversity on the dataset, so that the model can generalize well whatever the speaker's gender or speech language.

3. METHODOLOGY

Some steps will be taken in order to accomplish reasonable multilingual classification:

1. Data Preparation:
   (a) Load audio data for every language (*.opus* format);
   (b) Sample randomly and equally sized across every language dataset;
   (c) Join sampled datasets.
2. Transformations:
   (a) Resample and convert to stereo;
   (b) Resize to same length;
   (c) Compute Mel Spectogram features.
3. Modelling:
   (a) Divide into train and test sets (balanced by sex and spoken language);
   (b) Train different types of model;
   (c) Validate models with classification metrics;
   (d) Decide the best modelling technique for the chosen dataset.

Among the models, we will train traditional models (e.g. Random Forest, SVM, Naive Bayes, Logistic Regression,

etc.) and sofisticated models (e.g. Convolutional Neural Networks, Recurrent Neural Networks, etc.). Once found the best model, we'd like to state that Neural Network models tend to be more accurate and less prone to overfitting compared to traditional models.

4. LIBRARIES

We'll use the following Python libraries for the development:

- *librosa*: to load audio files and plot them;
- *scipy*: to manipulate signals;
- *matplotlib*: to plot signals and graphics;
- *sklearn*: to train traditional models;
- *PyTorch*: to create and train neural networks;
- *SHAP*: to give interpretative results about the models' outputs.

We're going to do every development on *Jupyter Notebook*, which are going to be available on a GitHub repository.

5. EXPECTED RESULTS

We expect to have a confusion matrix for each model and be able to compare the performance between them, so as to choose one that best learns our dataset features, and that generalizes the best for the test dataset.

As an additional feature, we would like to have some interpretation on the wave forms, in order to identify language specific features. For instance, it would be interesting to find formants and phonemes that describe Dutch speech in contrast to those that describe Portuguese speech.

6. REFERENCES

[1] *MLS: A Large-Scale Multilingual Dataset for Speech Research*, V. PRATAPM, Q. XU, A. SRIRAM, G. SYNNAEVE, R. COLLOBERT. ArXiv. 2020. abs/2012.03411.