Project Milestone Template

Maanusri Balasubramanian
University of Massachusetts
Amherst, MA
mbalasubrama@umass.edu

Sanjana Radhakrishna
University of Massachusetts
Amherst, MA
sanjanaradha@umass.edu

1. Introduction

We aim to classify the audio files as those spoken by a native speaker/non-native speaker and also rate the closeness of the pronunciation of a non-native speaker to a native speaker. We are given instances of native and non-native speakers reading out the same paragraph. So the overall plan is to use a convolutional neural network to classify these audio instances into those of native/non-native speakers. And also rate the closeness of these audio instances of non-native users to those of native speakers. This is basically like giving a score to a person based on how close their pronunciation is to a native speaker.

2. Problem Statement

The targeted problem is to classify the audio instance as that of a native/non-native speaker and also score the pronunciation of the speaker. A score closer to one, indicates that the speaker's accent is very much similar to that of a native speaker. Similarly a score closer to zero indicates that the speaker's accent is very different from that of a native speaker.

Dataset we are planning to use is from https://accent.gmu.edu/. The dataset consists of audio clips from a number of native and non-native, where each speaker repeats the same sentence. Here, we have the paragraph "Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station." repeated by a native and non-native speakers.

We could evaluate the scores produced by our model with reference to the word-level confidence scores google assigns to each of these user audio files.

3. Technical Approach

Our data is in the form of audio clips which we plan on transforming into images. We intend to transform our data into images using some features extraction methods like mel frequency cepstral coefficients (MFCC). These produced images can then be fed to a Convolutional Neural Network which is trained to classify instances into native and non-native speaker categories.

The pronunciation of the non-native speaker could then be scored (score of one being closer to a native speaker). During training, we will compare the image representation of the audio of one non-native speaker with each of the audio files of native speakers represented as images and conclude on a difference/distance measure. Using this distance as a base, we will be able to rate the pronunciation of a new test instance.

4. Intermediate/Preliminary Results

We have collected the data required for this task and are transforming it so that it can be directly used by a Convolutional Neural Network.

References

[1] Xia Zhan. A Convolutional Network-Based Intelligent Evaluation Algorithm for the Quality of Spoken English Pronunciation, 2022.

[2] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. Convolutional Neural Networks for Speech Recognition, 2014.