Proof of Concept

Bogdan Negru

May 2025

1 Model development

For this task, we have decided to reuse the Wav2Vec 2.0 model, developed by Facebook for the purpose of speech-to-text. We have trained this architecture on the English Accent Samples dataset, that contains 4,635 audio samples of the English language, from both native and non-native speakers.Due to imbalances in the data, each accent is limited at 105 samples(or lower in the case of accent with fewer)

The audio samples have been resampled to a standard 16 kHz, in order to achieve a consistent sample rate. Then we normalize the amplitude in order to obtain a relative constant level of loudness between samples. In the end, the samples have been padded to have the same length.

In training, we used the AdamW optimizer with a learning rate of 0.00005, for 15 iterations. The model has obtained a training accuracy of 95% in training and 88% in testing.

2 Application logic

In our script, we download the video from the provided url, extract the sound and the apply the same preprocessing done in the training of the model before obtaining the prediction.