



1. An expected further improvement of model performance would come from utilizing the full training dataset if we have sufficient RAM resources. Alternatively, a lower batch count during training and higher `gradient_accumulation_steps` can compensate for the RAM usage with more training time.
2. Data augmentation could be performed on the audio data, e.g. through low pass filtering, varying the tempo, or introducing artifacts like echo to artificially increase the training dataset.
3. The additional metadata that came with the datasets: age, gender, accent could be incorporated into the model by adding additional layers on top of the base model for the extra inputs. This information can help the model make better predictions.